Interpersonal differences in social dilemmas

Liebrand, Wilhelmus Bernardus Gerhardus

CHAPTER 2

A CLASSIFICATION OF SOCIAL DILEMMA GAMES

Social dilemmas have been defined by Dawes (1975) as situations in which: (1) each person has available a dominating strategy, i.e., one which yields the person the best payoff in ALL circumstances; and in which (2) the collective choice of dominating strategies results in a deficient outcome, that is, a result that is less preferred by all persons than the result which would have occurred if all had not chosen their dominating strategy. Dawes' requirement of a dominating strategy (1975, 1980) for each person does not appear to be crucial for considering a situation a social dilemma. What is critical is that a strategy can be chosen which ultimately results in an outcome which is deficient for all persons involved, and which nonetheless can be attractive since in SOME circumstances that strategy yields the best payoff for the person choosing that strategy.

Relaxing the dominance assumption enables one to evaluate more situations that have the psychological characteristics of social dilemmas. For this reason, in the present research a broader definition of the concept of social dilemma is formulated and employed. Here, a social dilemma is defined as a situation in which: (1) there is a strategy which yields the person the best payoff in at least one configuration of strategy choices, which has a negative impact on the interest of the other persons involved; and in which (2) the choice of that particular strategy by all persons results in a deficient outcome. The latter definition still has the advantage of being based on comparison of payoffs only WITHIN an individual (cf. Dawes, 1980). It differs from Dawes' definition in that instead of the dominant strategy which yields the best payoff in all circumstances, a strategy is employed which depending upon others' choices, might yield the best payoff; it matches Dawes' definition in that the choosing of that very strategy, does have negative consequences for others, and ultimately will result in a deficient outcome for all.

The parallelism between n-person games and social dilemmas has been noted frequently (Brechner, 1977; Dawes, Delay & Chaplin, 1974; Edney & Harper, 1978; Kelley & Grzelak, 1972; Kahan, 1974; Messick, 1973). Moreover, some n-person game classifications have been proposed (Dawes, 1980; Goehring & Kahan, 1976; Komorita, 1976; Weil, 1966). The present chapter attempts to extend this line of research, by investigating which game settings are capable of capturing the essence of the social dilemma structure as defined above. Drawing heavily on the work of Hamburger (1973, 1974), all 2-person 2-alternative games possessing social dilemma properties will be selected. Next, based on the 2-person social dilemma games, a classification of n-person social dilemma games is proposed. And finally, it will be shown that the present classification extends earlier classification schemes.

TWO-PERSON TWO-ALTERNATIVE GAMES WITH SOCIAL DILEMMA PROPERTIES

In the so called 2 x 2 games, each of two players has to choose, privately, one of two alternatives. The consequences to each player of each possible combination of choices, specified in the payoff matrix of the game, are known to both players in advance. The strategic properties of different types of 2 x 2 games can be analyzed by comparing the payoff matrices of the games.

The number of different 2 x 2 games which can be constructed is infinite. Therefore, following Rapoport and Guyer (1966), some restrictions are introduced. First, the present analysis is based upon the preference ordering of the four payoffs as they appear to the one and to the other player. Second it is supposed that each player has a strict preference ordering of the four possible payoffs. Given the two restrictions, there still are 4! x 4! = 576 ways to fill up the payoff matrix. After eliminating games which are invariant up to an interchange of rows and/or columns and/or players, there are 78 nonequivalent 2 x 2 games possible, (Rapoport & Guyer, 1966). In this section, a subclass of these nonequivalent 2 x 2 games will be considered. This subclass consists of games which are both (a) symmetric, - - that is, games which "look the same" to both players (Harris, 1969, p. 139), and (b) which also possess the two social dilemma properties described previously. As is shown below, there are only three 2 x 2 game formats conforming to these requirements.

Setting aside the attractiveness restriction for a moment, the first social dilemma property is tantamount to the availability of a strategy having negative consequences for the other person. This property reduces the number of potentially relevant games to those games where both players have a most-threatening strategy. A strategy is called most-threatening if a rational Player 1 prefers Player 2 not to choose that strategy, irrespective of Player 1's choice. In such a case Player 2 has a most-threatening strategy (Hamburger, 1973;1974). Following Hamburger's (1974) method of proof, it is easily seen that there are exactly six symmetric games where each player has the choice between strategy A and the most-threatening strategy B.

First, a player's strict preference ordering of the four payoffs is labelled 4, 3, 2, 1 in decreasing order of preference. Then the payoffs to each player can be distributed into the four cells of the payoff matrix in such a way that the outcome to Player 1 is always stated first in a cell. Next, without reducing the number of 78 nonequivalent 2 x 2 games, it is possible to put Player 1's most-threatening strategy in the second row, and Player 2's most-threatening strategy in the second column. Given this strategy configuration, it follows that the payoffs to Player 1 in the left column must be

Table 1

Six symmetric 2 x 2 games; alternative B is each player's most-threatening strategy; entries are preference orderings: (4) best possible outcome, (1) worst possible outcome; the first entry refers to Player 1, the second to Player 2; Player 1 is row player, Player 2 is column player.

Matrix 1

	A	B
A	3,3	1,4
B	4,1	2,2

Matrix 2

	A	B
A	3,3	2,4
B	4,2	1,1

Matrix 3

	A	B
A	4,4	1,3
B	3,1	2,2

Matrix 4

	A	B
A	2,2	1,4
B	4,1	3,3

Matrix 5

	A	B
A	4,4	2,3
B	3,2	1,1

Matrix 6

	A	B
A	4,4	3,2
B	2,3	1,1

higher than his or her' payoffs in the right column. Therefore, payoff 4 to Player 1 can appear only in two outcome cells, the other payoff to Player 1 in the same row can be either 3, 2, or 1. After assigning two payoffs in this way, the other payoffs to Player 1 are uniquely determined. Consequently, a total of 6 different payoff configurations to Player 1 can be distinguished. Finally, out of the 6 (Player 1) x 6 (Player 2) = 36 payoff matrices, only the six symmetric matrices have to be considered. These six matrices are shown in Table 1.

Having a most-threatening strategy thus reduces the number of 2 x 2 games to be considered to six. However, Social Dilemma Property 1 furthermore implies that this most-threatening strategy has to be attractive to at least one rational player. Here, strategy B is considered to be potentially attractive if, for at least one pair of strategy-choices, strategy B yields the best payoff to at least one player. In other words, all the payoff matrices in which strategy A is a dominant strategy, have to be eliminated. In Table 1 this affects Matrices 5 and 6.

Finally, Social Dilemma Property 2 eliminates the game depicted in Matrix 4. In 2-person settings, Property 2 states that both players are better off if both choose A than if both choose B. Since the pair of outcomes resulting from both players choosing B is higher than the outcomes resulting from an A choice by both players, the restriction imposed by Property 2 is not fulfilled in Matrix 4.

Out of the three symmetric 2 x 2 games with social dilemma properties, Matrix 1 and Matrix 2 have undergone extensive experimental investigation. Matrix 1 is the well-known Prisoner's Dilemma Game (Luce & Raiffa, 1957); Matrix 2 is known as Chicken (Kahn, 1965). The game depicted in Matrix 3, labelled by Rapoport and Guyer (1966, p. 209) as trivial no-conflict game, is not very well known. For present purposes Matrix 3 will be labelled the "Trust Game".

THE DECISIONAL STRUCTURE OF THE PRISONER'S, THE CHICKEN AND THE TRUST SOCIAL DILEMMA

The following anecdote, taken from Luce and Raiffa (1957) illustrates both the name and the social dilemma properties of the Prisoner's Dilemma. Two individuals, accused of robbing a bank, are taken into custody and separated. The district attorney, unable to prove that they are guilty, confronts each prisoner with two alternatives: either confess to the crime (Alternative B), or not confess to it (Alternative A). If both suspects confess, each will receive a 5-years' sentence. If neither suspect confesses, both will be convicted on some minor charge and receive a 1-month sentence. If one confesses and the other does not, the suspect who does not confess will receive a 10-years' sentence while the other will be set free. The consequences associated with the four possible combinations of choices are such that they result in the preference orderings depicted in Matrix 1 (Table 1). The ordering for each prisoner is strict, ranging from 1 (worst outcome: 10-years sentence) to 4 (best outcome: free). As appears from the preference orderings it is to each prisoner's advantage to confess, regardless of the other's choice. However, if both prisoners act in their own interest and confess, they both end up in a worse position (5-years sentence) than in case they do not confess (1-month sentence). The Prisoner's Dilemma is an example of a mixed-motive game, there is a motive to cooperate (A: not confess), and there is the incompatible motive to defect (B: confess). The specific ordering of payoffs as depicted in Matrix 1 results in two important properties of the Prisoner's Dilemma Game. First, each player has a dominating strategy. Second, if both players choose their dominant strategy, which is prescibed by the principle to maximize the payoff or the principle to maximize the minimum payoff (maximin), a deficient outcome results.

The term "Chicken" (Matrix 2) applies to a game which is according to Broeze (Note 2) popular among American teenagers. Two young Americans are sitting in a fast driving car. The driver takes his hands off the steering wheel. The "chicken" is the one first taking the steering wheel (A-alternative), giving the other thereby the highest payoff. However, the joint disaster, or the worst outcome for both results from a joint refusal to steer again.

In contrast to the Prisoner's Dilemma, in Chicken the most-threatening strategy is not a dominating strategy. In Chicken the most-threatening strategy B is the strategy to be selected by players trying to get the highest payoff (maximax principle). Both in Prisoner's Dilemma and in Chicken a double B-choice results in a deficient outcome. Only in Chicken, however, is this deficient outcome the worst possible outcome.

Finally, the following anecdote illustrate the decisional structure of the Trust Game. In order to decide which one is the best long-distance runner, two athletes plan to run a marathon. Both prefer an honest race to a race in which one or both of them are using a drug (alternative B). However, given that the other is going to use dope, each prefers to use dope so as to avoid being at a disadvantage in the race.

The most-threatening strategy B (i.e., to use drugs) is not a dominating strategy in the Trust Game. Choosing B, however, clearly is prescribed by the well-known strategy to maximize the minimum payoff (maximin principle). Doing so results in a deficient outcome for both players.

FROM MATRICES TO GRAPHS

To represent payoffs in three- or more-person games in the same way as was done for 2 x 2 games, a three or more dimensional matrix would be required. To avoid these cumbersome matrices, a graphical representation is used in Figure 1 to illustrate the three games discussed above.

Again the B-alternative corresponds to the most-threatening strategy, the A-alternative to the common interest strategy. The number of players choosing alternative A is depicted on the horizontal axis; the two payoff graphs for each game refer to the payoffs for a player choosing either A or B, given a particular total number of A-choices. The correspondence between Matrices 1 to 3 (Table 1) and Graphs 1 to 3 (Figure 1) can be seen by comparing the payoff orderings. On the one hand the payoff orderings are presented by the matrix cell entries, on the other hand they are presented by the end points of the graphs. For example, in Prisoner's Dilemma (Matrix 1; Graph 1), the payoff graph for choosing B includes the highest payoff (4), and lies in all circumstances (0, 1, or 2 A-choices) above the graph for choosing A. However, the end point representing the payoff for both, in case both had chosen A (3), is above the left end point of graph B (2).

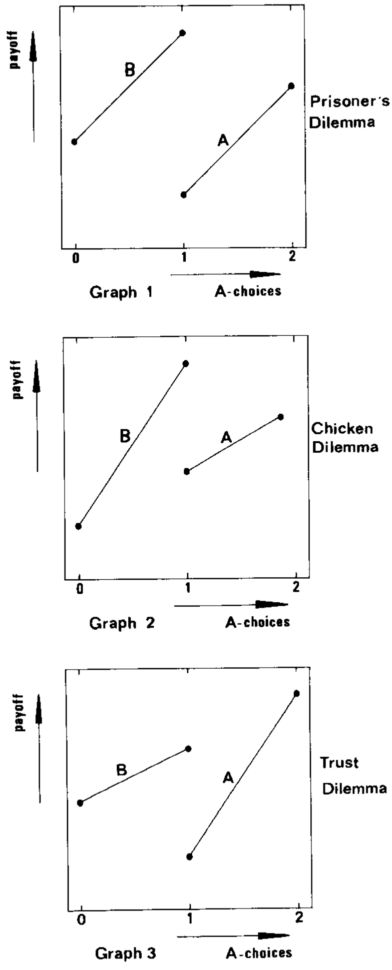

Figure 1

Three 2x2 games possessing social dilemma properties; alternative B is the most-threatening strategy

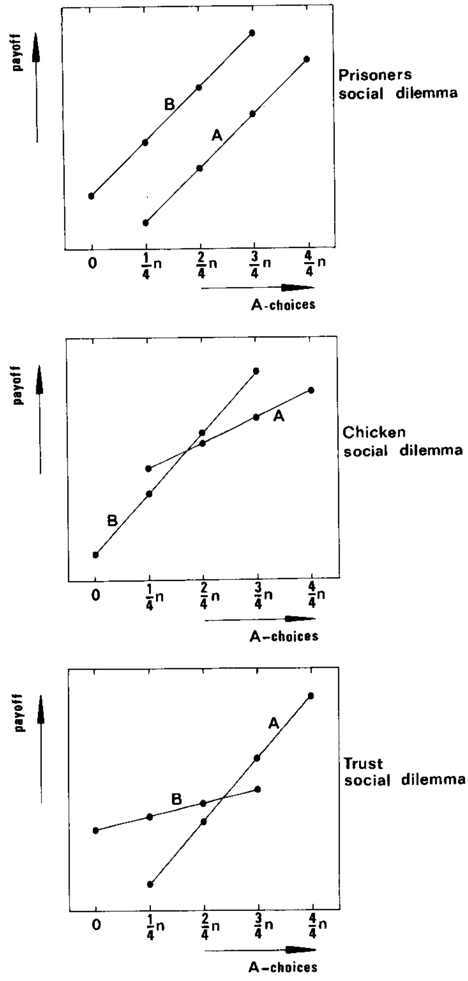

Figure 2

Three n-person social dilemma games; alternative B is the most-threatening strategy

FROM TWO-PERSON TO N-PERSON GAMES

In turning from two-person to three- or more-person games, some distinctive characteristics are introduced (Dawes, 1980). First, in two-person games there is one opponent, choosing either A or B, so that each player knows with certainty how the other has behaved from the payoff received. In n-person games this identification is impossible. As long as not all the other players make the same choice, all personal choices remain secret. Second, the influence of one individual's choice on the other's payoff, called the "externality" (Buchanan, 1971; Schelling, 1973), is spread out over a considerable number of players. In contrast, in two-person games, negative or positive externalities are focussed, they directly punish or reward the other player.

The increased anonimity and the greater spread of externalities do not alter the decision structure in the three types of social dilemmas. They may lower the threshold for choosing the most-threatening strategy - - constituting thereby, compared to the 2-person version, an even more threatening situation. Thus, regarding the parallelism between the decision structure of the 2-person and the n-person games, the three 2 x 2 games with social dilemma properties provide a useful classification for n-person social dilemma games. In the following three real-life examples are used to illustrate the payoff graphs of the three types of n-person social dilemma games. In order to simplify the analysis, the choices of subgroups consisting of 1/4 n persons are considered to be equal. Consequently, the payoff graphs are defined in case 4/4 n, 3/4 n, 2/4 n, 1/4 n, or 0 persons have chosen alternative A.

Prisoner's Social Dilemma. The decision to pollute may be described by the Prisoner's Social Dilemma payoff graphs in Figure 2 (cf. Dawes et al., 1974; Goehring & Kahan, 1976). Pollution problems can be found at various levels of decision making, ranging from individuals to nations. For example, at the industrial level, no matter what other chemical industries may do to get rid of their chemical waste, it is cheaper to have the waste dumped at some rubbish-dump, or alternatively, in the ocean, than to take care of an adequate decomposition of the waste. The ultimate long-term consequences of this selfish act have to be shared by all individuals. At the individual level the slogan "every litter bit hurts" nicely reflects the negative externalities accompanying the decision to pollute (alternative B). Though all individuals would like to avoid the long-term negative consequences, it remains cheaper and simpler for them to keep polluting as anonymous individuals. The payoff graph for polluting lies for its entire length above the graph indicating the payoff for not polluting.

Chicken Social Dilemma. In the process of deciding whether to go by bike (alternative A) or by car (alternative B), an individual trying to get to work as fast as possible is facing a Chicken Social Dilemma. In this type of social dilemma the payoff for choosing either A or B, strongly depends upon the number of others deciding to go by bike (A). If hardly anybody goes by bike, there will be many cars on the road and consequently there will be traffic jams. That being so, our decision maker is better off going by bike than by car. In Graph 2 (Figure 2) this situation is reflected to the left of the intersection where the payoff graph for alternative A lies above the one for alternative B. Instead of going by bike, the same person is better off going by car if many people decide to go by bike: to the right of the intersection the payoff for alternative A becomes less and less attractive. In case everybody decides to go by car the worst possible outcome for all occurs. The accumulation of the negative externalities is then expressed in congestion and polluted air. In that case, each person would prefer a situation in which neither of them would use a car.

Trust Social Dilemma. There are times at which a food supply is not excessive but sufficient. Any initiation of hoarding, however, generates a Trust Social Dilemma. Clearly the highest payoff results from no hoarding at all (alternative A). Not hoarding food, for example milk, provides no additional costs for preservation while there is enough milk available in the stores. If only a small number of persons are keeping a lot of milk in reserve, one is better off not hoarding milk. In Graph 3 (Figure 2) this is reflected by the payoff graphs to the right of the intersection. But if the number of persons hoarding milk increases, the attractiveness of one's own hoarding increases. At the end, thanks to the massive hoarding there will be no more milk available in the stores. Consequently, in that case the worst possible payoff accrues to the person who had chosen alternative A.

RELATIONSHIP WITH OTHER CLASSIFICATION SCHEMES

Most n-person game classifications consider the Prisoner's Dilemmas. Typically, a n-person 2-alternative Prisoner's Dilemma ($n > 2$) is defined by:

$$B(j-1) > A(j) \qquad j = 1, 2, \ldots, n \tag{1}$$

$$A(n) > B(0) \tag{2}$$

where the index within parentheses represents the number of A-choices; $B(j)$ the payoff to each player choosing B, given the total of j A-choices; and $A(j)$ the payoff for choosing A in that case.

Weil (1966) suggests a categorization based on the algebraic sign of ($A(j) - A(j-1)$) on the one hand, and ($B(j) - B(j-1)$) on the other hand. Together with the assumption that all members of the set (A(j) - A(j-1)) and the set (B(j) - B(j-1)) are alike with respect to algebraic sign, and the restriction that j is not equal to n, four cases can be distinguished: POSITIVE-POSITIVE, POSITIVE-NEGATIVE, NEGATIVE-POSITIVE and NEGATIVE-NEGATIVE. Both Weil (1966) and Goehring and Kahan (1976) consider the POSITIVE-POSITIVE case the one in which most applications to the real world can be found.

Goehring and Kahan further subdivide Weil's POSITIVE-POSITIVE case, or those games having payoff functions increasing with the number of A-choices, into three types of Prisoner's Dilemma games. Type 1 and Type 2 consists of those games having payoff functions whereby (B(j-1) - A(j)) increases or decreases with the total number of A-choices, respectively. Type 3 in Goehring and Kahan's classification consists of those games having linear parallel payoff functions. They designate this type of game the "uniform" Prisoner's Dilemma.

Dawes does make a still further subdivision of the uniform Prisoner's Dilemmas. Following Hamburger (1973), he distinguishes "take some" and "give some" games. The difference between the two games lies in the procedures used: in take some one can take some from others, and in give some one can contribute to a common good. In addition to these uniform games, Dawes' classification of social dilemma games consists of "variable games", or games which because of their complicated rules and regulations defy a simple mathematical description of the payoff configuration. (e.g., Rubenstein, Watzke, Doctor & Dana, 1975).

All the above classifications are based on Prisoner's Dilemmas conforming to specific requirements. In addition to these classifications, one more model of n-person games has been proposed (Komorita, 1976). Komorita (p. 358) defines n-person dilemmas rather unconventionally. He defines the essential conditions as: "1. each of n persons has two choices, cooperative (A) or competitive (B); 2. the outcomes for both choices increase monotonically with the proportion of people who make the cooperative choice; 3. the competitive choice always yields a higher outcome than the cooperative choice; and 4. the outcome if everyone makes a cooperative choice is greater than the outcome if everyone makes a competitive choice." Next, Komorita (p. 359,360) states that "the essential condition that the B-choice dominates the A-choice implies that . . . B(j) > A(j)". In defining the concept of dominance in this way, it is possible that B(j-1) < A(j), that is becoming the (j-1) + 1st A-chooser yields a higher payoff than the B choice would afford in that case. This very payoff configuration does not satisfy the Prisoner's Dilemma requirements. On the other hand, given a Prisoner's Dilemma Komorita's condition B(j) > A(j) is true. Consequently, Komorita's model captures more types of games than just the Prisoner's Dilemma. In addition Komorita proposes an index of cooperation (K*) based on Rapoport's (1967) index for the two-person Prisoner's Dilemma. K* is defined by Komorita as:

$$A(n) - B(0) / O(max) - O(min)$$

where O(Max) and O(min) denote the maximum and minimum possible outcomes. The index K* then serves to distinguish different types of n-person dilemma games. However, K* can take the same value given two different types of n-person games. For example, consider the Chicken social dilemma and the Trust social Dilemma depicted in Figure 1. If the payoff graphs represent numerical values ranging from 4 to 1, K* equals (3-1)/(4-1) for the Chicken Dilemma and (4-2)/(4-1) for the Trust Dilemma. Taking into account the insensitiveness of K* for different payoff structures, in the present research Komorita's model is considered not a useful model for classifying n-person dilemma games.

As was stated previously, the present classification extends the above classifications in that it consists of three different types of social dilemma games, based on an exhaustive set of 2 x 2 games. It consists of the Prisoner's Social Dilemma, of which no further subdivision is provided, the Chicken Social Dilemma and the Trust Social

Dilemma. Until now empirical research has been focussed on the Prisoner's Social Dilemma (Dawes et al., 1977; Caldwell, 1976; Kelley & Grzelak, 1972) and the Chicken Social Dilemma (Meux, 1973). However, there seems to be no apparent reason to exclude the Trust Social Dilemma from n-person game research.

DISCUSSION

Social dilemmas were defined as situations in which by the very act of choosing a strategy with negative externalities, the ultimate outcome can be called deficient. Starting from Rapoport and Guyer's (1966) taxonomy of nonequivalent 2 x 2 games, it has been shown that exactly three of these games possess the social dilemma properties as defined. Next, it appeared that the decision structure underlying different real-life situations can be properly captured by the n-person generalizations of the three 2 x 2 social dilemma games.

The payoff configuration for the three types of social dilemmas provides insight into the reasons for behaving in such a way that a deficient collective outcome results. Not choosing the dominant strategy with negative externalities in a Prisoner's Social Dilemma is called an irrational way of behaving. Two other "rational" ways of behaving or selection principles can be distinguished (Hamburger, 1979, p. 42). The principle of maximizing the maximum payoff and the principle of maximizing the minimum payoff both prescribe the decision to choose the strategy with negative externalities, in a Chicken- and a Trust Social Dilemma, respectively. So, in all three types of social dilemmas the behavior which is not in the service of the common interest is prescribed by the above selection principles. It follows that the most likely outcome is the deficient outcome. Hence, the development of solutions to avoid the deficient outcomes can be called the most important task of the social dilemma paradigm.

The three types of n-person social dilemma games provide a promising research tool for the development of such solutions. The social dilemma mechanism can be captured in laboratory analogues. It is a common observation, that in such instances subjects do take the decision task extremely serious (Bonacich, 1976; Dawes et al., 1977). Dawes et al. (p. 7), for example, write:

> One of the most significant aspects of this study, however, did not show up in the data analysis. It is the extreme seriousness with which the subjects take the problems. Comments such as, "If you defect on the rest of us, you're going to live with it the rest of your life," were not at all uncommon. Nor was it unusual for people to wish to leave the experimental building by the back door, to claim that they did not wish to see the "sons of bitches" who double-crossed them, to become extremely angry at other subjects, or to become tearful.

However, in employing games as a research tool one significant problem is the way in which the payoff matrix as presented by the experimenter is actually experienced by the subjects. Kelley and Thibaut (1978) point out that the outcome matrix presented by an experimenter, which may be described as the "given" matrix, may not be the one on which the decisions of the actors are based. Rather actors may transform the outcomes in a given matrix into utilities according to the personal values they place on the alternative outcome distributions their choices would afford themselves and other persons (Harris, 1969; Kelley & Thibaut, 1978; McClintock, 1972). Kelley and Thibaut describe this process as one of moving from a "given" to an "effective" matrix.

In the present chapter it was assumed throughout that each player has a strict preference ordering of the outcomes, or alternatively, matrix cell entries represent player's utilities. Given these utilities each person faces the same type of dilemma. However, it follows that differentially transforming the numerical outcomes in the cells of the given matrix affects the structure of the game accordingly. For example, suppose that the numerical outcomes in Matrix 1 (Table 1) represent money instead of utilities. The Prisoner's Dilemma structure, evident for a person focussed on payoffs to self, is then absent for a person trying to maximize the other's payoff (altruism). A fortiori, such a dollar representation of either Matrix 1, Matrix 2, or Matrix 3 (Table 1), would not generate a social dilemma at all, if all persons were more concerned with the payoff to others than with their own payoff. Such a case, however, is considered strictly hypothetical here. In the present research, subjects are confronted with a decision task in which at least for persons only concerned with their own payoff, the social dilemma structure exists. In Chapter 3 this experimental decision task will be described in more detail and a model will be proposed to capture the process of differentially transforming the given matrix. In addition, specific predictions derived from the proposed model will be investigated in two experiments under different experimental conditions.